DOCUMENT RESUME

ED 446 994 SE 064 299

AUTHOR Obiekwe, Jerry C.
TITLE A Comparative Analysis of Mathematics Placement Test Using Item Response Theory.
PUB DATE 2000-10-00
NOTE 14p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 25-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Difficulty Level; Elementary Secondary Education; *Item Response Theory; Mathematics Achievement; *Mathematics Instruction; Mathematics Tests; Measures (Individuals); Student Placement; Test Items; Testing

ABSTRACT

The first purpose of this study is to do a comparative analysis in terms of the item difficulty and discrimination index between Mathematics Placement Test Form A and G. The results seem to suggest that items on both forms of the placement test are equally difficult. However, items on form A appear to have more discrimination power than items on form G. Perhaps form A should be used more frequently than form G in making math placement decisions. The second purpose was to determine a subset of items from form A that can better predict students' success in mathematics classes. The third purpose was to determine a subset of items from form G that can better predict students' success in mathematics classes. Those items whose item discrimination power were greater than 1 were the reliable items. The identified reliable items can be used exclusively or they can be weighted more and then used in conjunction with the less reliable items in placing students into math classes. This process, perhaps, may guard against placing students into classes where they do not belong. (Contains 17 references.) (YDS)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 446 994

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. Obiekwe
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A Comparative Analysis of Mathematics Placement Test Using Item Response Theory

Jerry C. Obiekwe
The University of Akron-Wayne College

Paper presented at the annual conference of the Mid-Western Educational Research Association. October 25-28, 2000, Chicago, IL

SE 061 299

BEST COPY AVAILABLE

A Comparative Analysis of Mathematics Placement Test Using Item Response Theory

Mathematics Placement Test

Mathematics Placement Test Form A and Form G are mathematics tests that are used to place students in mathematics courses depending on their mathematics proficiency. At this college, Form A and Form G are the most commonly used in placing students in math courses. Form A or Form G contains 40 multiple choice items, hence 40 possible points.

Often times students' placement scores were not good predictors of their success in their placed math courses. One possible reason could be that their placement score is a composite score resulting from both reliable items and less reliable items. Since all items are weighted equally, students whose scores on the placement test primarily came from the less reliable items are more likely to be placed in a math class that they may have difficulty succeeding. Perhaps one way to guard against this type of misplacement is to give more weight to the reliable items when calculating the placement scores, and in turn raise the cut off scores for the placement of students in math courses. Alternatively, reliable items on the placement test could become the sole determinant in placing students into various levels of math courses.

Reliable items on a placement test are items that can discriminate between high ability students and low ability students. The reverse is the case with less reliable items. Therefore, scores on reliable items on a math placement test can be better predictors of success in math classes.

Objective of the Study

The purpose of this study is to do a comparative analysis in terms of the item difficulty and discrimination index between Mathematics Placement Test Form A and G. The second purpose is to determine a subset of items from form A that can better predict students' success in mathematics classes. The third purpose is to determine a subset of items from form G that can better predict students' success in mathematics classes.

Background Information on Item Response Theory.

Item response theory (IRT) is a mathematical model that relates the probability of answering an item on a test correctly to the ability of the student, the difficulty of the item, and the discrimination of the item (see equation 1). These three parameters, student ability, item difficulty, and the item discrimination, are unknown and will be inferred from the student responses (Hambleton, Swaminathan & Rogers, 1991; Hulin, Drasgow & Parsons, 1983; Lord, 1980).

$$P_i(\theta) = c_i + (1 - c_i)\frac{\exp Da_i(\theta - b_i)}{1 + \exp Da_i(\theta - b_i)} \qquad i = 1,2,.....n \qquad (1)$$

Equation 1 is the three parameter version of the item response theory (Birnbaum, 1968), where $P_i(\theta)$ is the probability of answering item i correctly, θ represents the ability of the student or the latent trait, b_i is the difficulty of item i, a_i is the discrimination index of item i, c_i is the lower asymptote of the item characteristic curve which corresponds to the probability of correct response to item i of the examinees with low θ, and D is a scaling constant and is usually set at 1.7.

Equation 1 collapses to two parameter model of IRT (Lord, 1952) if $c_i = 0$ (see equation 2).

$$P_i(\theta) = \frac{\exp Da_i(\theta - b_i)}{1 + \exp Da_i(\theta - b_i)} \qquad i = 1,2,.....n \qquad (2)$$

where $P_i(\theta)$, θ, b_i, a_i, D are the same as in equation 1.

Equation 1 reduces to one parameter model of IRT if $c_i = 0$, a_i=1, D=1(see equation 3).

$$P_i(\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)} \qquad i = 1,2,.....n \qquad (3)$$

Equation 3 is often referred to as the Rasch model in honor of its developer (Rasch, 1966 & Rasch, 1980; Gustafsson, 1980; Harris, 1989).

Test Administration

This study analyzed the item responses of 288 freshmen who took Mathematics Placement Test Form A in the fall of 1999. Likewise, the study analyzed item responses of 280 freshmen who took Mathematics Placement Test form G in the spring of 2000. The analysis provided insight into the item difficulty and the item discrimination and the reliability of the test.

Reliability of the Placement Tests Form A and Form G

The reliability of the test will be evaluated by the internal consistency of the items which gives the lower bound of the actual reliability (Novick & Lewis, 1967). By using the SPSS package, the internal consistency of Placement Test form A was found to be .86, while Placement Test Form G was also found to be .86 also.

Assumptions

The IRT models assume that a single dominant factor or ability accounts for examinee performance on the test. This assumption is called unidimensionality. The assumption cannot be strictly met since there are other intervening factors that may affect test performance. Essentially, with regard to this study, what this assumption is saying is that if other intervening factors that may affect test performance are held constant, then the only factor responsible for examinee performance is mathematics proficiency. The second assumption, which is related to unidimensionality, is local independence. Local independence is the concept that the examinee's performance is only related to the latent trait. When the assumption of unidimensionality is met, so also is local independence (Lord & Novick, 1968).

Unidimensionality of Placement Test Form A

In order to ascertain whether the assumption of unidimensionality is met in this study with regard to Placement Test Form A, two different methods were applied. In the first method, the item responses were submitted to tetrachoric factor analysis using McroFact computer program. Two factors were extracted(see Table 1). The first factor explained 23.3% of the total variance, while the second factor explained 2.3%. The second method was to submit a one factor model to confirmatory factor analysis via EQS. The comparative Fit Index (CFI) was found to be 0.80. Based on the results of these two methods, it is reasonable to assume that the requirement of unidimensionality was met, since Reckase (1979) suggests that at least 20% of the test variance be explained by the first factor, and Bentler(1992) wants the CFI to be greater than .90.

Unidimensionality of Placement Test Form G

Similarly, in order to ascertain whether the assumption of unidimensionality is met in this study with regard to Placement Test Form G, two different methods were applied. In the first method, the item responses were submitted to tetrachoric factor analysis using McroFact computer program. Two factors were extracted(see Table 2). The first factor explained 24.1% of the total variance, while the second factor explained 1.9%. The second method was to submit a one factor model to confirmatory factor analysis via EQS. The comparative Fit Index (CFI) was found to be 0.85. Based on the results of these two methods, it is reasonable to assume that the requirement of unidimensionality was met, since Reckase (1979) suggests that at least 20% of the test variance be explained by the first factor, and Bentler(1992) wants the CFI to be greater than .90.

Checking the model fit for Placement Test Form A and Form G

It is required that the fit of the IRT model to the data be assessed before their application. The fit to a set of test data implies that the model can explain the data. It also means that the ability estimates obtained from different sets of test items will be the same, while the

item parameter estimates derived from different groups of examinees will also be the same. This characteristic of IRT models when the data fit the model is called the property of invariance.

An important question is whether the three parameter model will provide a better fit than the two parameter model. The BILOG program provides likelihood statistics at the end of each cycle of iteration. Therefore by comparing the likelihood ratio chi-square of the two parameter and the three parameter models, with the degrees of freedom equal to the difference in the number of parameter accounted for times the number of items, it can be determined which model provides a better fit. This is tantamount to testing the hypothesis whether an additional parameter does make a difference (Mislevy & Bock, 1990).

At the final cycle of iteration for item responses of Placement Test Form A, the likelihood ratio chi-square of the two parameter model should be greater than that of the three parameter model (see Table 3). The difference between the two fit statistics was found to be significant (see Table 4). The three parameter model did provide a better fit than the two parameter model.

Similarly, at the final cycle of iteration for item responses of Placement Test Form G, the likelihood ratio chi-square of the two parameter model should be greater than that of the three parameter model (see Table 5). The difference between the two fit statistics was not significant (see Table 6). The three parameter model did not provide a better fit than the two parameter model. Consequently, two parameter model was used to derive the item difficulty, and the item discrimination index for Placement Test Form G.

The whole test and item fit statistics for the two and three parameter models were provided by the use of the same computer program BILOG. This program reported the chi-square statistics for the fit of each item, and the whole test (see Table 7). Of the 40 items in Form A none was misfitted because all the reported probability values were greater than the critical probability level of .01. The same was also applicable to Form G.

Analysis of Table 8: Ability, Item Difficulty and Discrimination Idex

The BILOG program provided estimates for the item difficulty and the discrimination parameter for both Placement Test Form A and G as shown in Table 7. The first column represents the discrimination indices for Placement Test Form A. The second column represents the discrimination indices for Placement Test Form G. The third column represents the item difficulty for Placement Test Form A , while the fourth column represents the item difficulty for Placement Test Form G.
The average ability parameter for Form A was –0.022(SD = 1.092), while the mean ability parameter for Form G was 0.005(SD = 1.125). Clearly, the average mathematics proficiency of students from the both samples was equal since there was no significant difference between the means ($t(288)=.29$, $p>.05$). The equality of the mathematics proficiency of both samples allows for the comparative analysis of the item difficulty and the discrimination indices in both samples.

The mean of the item difficulty of the Placement Test Form A was 0.638 (SD = 1.466) while the mean of the item difficulty of the Placement Test Form G was 0.529 (SD = 1.620). The difference between the two means was not significant, $t(288) = .84$, $p>.05$. This may suggest that on average the items on Placement Test Form A and G are equally difficult.

The mean of the discrimination index for Placement Test Form A was 1.228(SD = .509), while the mean of the discrimination index for Placement Test Form G was 1.000(SD = .491). The difference between the means was significant, $t(288) = 5.454$, $p<.05$. This may suggest that on average items in Placement Test Form A appear to have more discrimination power than items in Placement Test Form G.

Items whose discrimination indices are greater than 1 with regard to Placement Test Form A are as follows: 3, 4, 9, 11, 12, 13, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 30, 31, 32, 33, 34, 36, 37, 38, 39. Items whose discrimination indices are greater than 1 with regard to Placement Test Form G are as follows: 3, 4, 9, 10, 11, 12, 13, 17, 18, 19, 21, 23, 25, 28, 33. According to Mislevy & Bock (1990), items whose discrimination indices are greater than one are more reliable than items with discrimination indices less than one but greater than zero.

Summary and Conclusions

The first purpose of this study is to do a comparative analysis in terms of the item difficulty and discrimination index between Mathematics Placement Test Form A and G. The results seem to suggest that items on both forms of the placement tests are equally difficult. However, items on form A appear to have more discrimination power than form G. Perhaps form A should be used more frequently than form G in making math placement decisions in this college.

The second purpose is to determine a subset of items from form A that can better predict students' success in mathematics classes. Those subset of items were found to be as follows: 3, 4, 9, 11, 12, 13, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 30, 31, 32, 33, 34, 36, 37, 38, 39 because their discrimination indices were greater than 1 hence more reliable than other item in the test. The third purpose is to determine a subset of items from form G that can better predict students' success in mathematics classes. Items 3, 4, 9, 10, 11, 12, 13, 17, 18, 19, 21, 23, 25, 28, 33 were found to be more reliable than other items on the test.

These items can be used exclusively or they can be weighted more and then used in conjunction with the less reliable items in placing students into math classes. Alternatively, each item can be weighted according to their item discrimination index. This process, perhaps, may guard against placing students into classes where they do not belong.

Table 1. Tetrachoric Factor Analysis of placement test form A

Factor 1	Factor 2
-0.951	0.013
-0.826	0.130
-0.781	0.101
-0.903	0.142
-0.853	0.266
-0.769	0.238
-0.663	0.298
-0.828	0.098
-0.922	0.00
-0.694	0.341
-0.804	0.186
-0.897	0.054
-0.762	0.385
-0.762	0.339
-0.663	0.425
-0.855	0.239
-0.827	-0.087
-0.859	-0.110
-0.708	-0.016
-0.891	0.077
-0.811	-0.032
-0.855	-0.022
-0.614	-0.513
-0.801	0.017
-0.775	-0.034
-0.665	-0.484
-0.787	-0.148
-0.751	0.037
-0.690	-0.409
-0.692	-0.225
-0.619	-0.503
-0.733	-0.249
-0.488	-0.301
-0.558	0.030
-0.635	-0.261
-0.569	-0.239
-0.787	-0.149
-0.660	-0.192
-0.566	-0.109
-0.923	0.079

Table 2. Tetrachoric Factor Analysis of placement test form G

Factor 1	Factor 2
-0.944	0.149
-0.840	-0.090
-0.835	0.019
-0.939	0.130
-0.855	0.120
-0.779	0.422
-0.668	0.231
-0.863	0.090
-0.961	0.116
-0.830	-0.016
-0.842	0.076
-0.877	-0.026
-0.561	0.336
-0.812	0.182
-0.681	0.258
-0.864	-0.012
-0.863	-0.105
-0.887	0.018
-0.773	-0.158
-0.896	-0.069
-0.706	0.133
-0.850	-0.171
-0.618	0.297
-0.835	-0.155
-0.736	-0.350
-0.655	0.036
-0.868	-0.114
-0.730	-0.234
-0.745	0.092
-0.686	-0.192
-0.657	-0.199
-0.787	-0.355
-0.692	-0.379
-0.708	-0.027
-0.617	0.145
-0.458	-0.576
-0.827	-0.126
-0.606	0.185
-0.612	-0.168
-0.731	0.374

Table 3
The likelihood ratio chi-square for two and three parameter models for Mathematics Placement Test Form A

Test	Two Parameter	Three Parameter
Placement Test Form A	12346.7313	12283.1015

Table 4
The difference between two and three parameter fit statistics for Mathematics Placement Test Form A

Two parameter – Three parameter	
63.6298*	

*p<.05

Table 5
The likelihood ratio chi-square for two and three parameter models for Mathematics Placement Test Form G

Test	Two Parameter	Three Parameter
Placement Test Form G	11958.7660	11920.7831

Table 6
The difference between two and three parameter fit statistics for Mathematics Placement Test form G

Two parameter – Three parameter	
37.9829	

p>.05

Table 7. Item fit statistics for Math Placement Test form A and G

Item	Chi-square	DF	Prob-Value	Chi-square*	DF*	Prob-value*
1	11.5	6.0	0.0738	3.9	9.0	0.9199
2	6.0	4.0	0.1983	5.7	4.0	0.2222
3	4.3	8.0	0.8317	9.2	7.0	0.2369
4	6.2	8.0	0.6289	5.2	6.0	0.5235
5	5.6	6.0	0.4716	1.1	6.0	0.9778
6	6.8	8.0	0.5627	7.4	8.0	0.4952
7	11.1	9.0	0.2662	9.2	8.0	0.3275
8	10.2	9.0	0.3369	5.4	9.0	0.7956
9	15.7	8.0	0.0462	4.8	6.0	0.5744
10	3.0	6.0	0.8106	4.5	3.0	0.2127
11	14.6	9.0	0.1034	4.9	7.0	0.6725
12	10.7	8.0	0.2187	9.5	6.0	0.1475
13	10.7	5.0	0.0572	6.0	5.0	0.3050
14	8.3	9.0	0.5055	7.4	9.0	0.5980
15	8.7	8.0	0.3672	11.9	8.0	0.1541
16	10.6	9.0	0.3050	8.1	9.0	0.5205
17	8.3	6.0	0.2166	7.9	6.0	0.2454
18	5.9	6.0	0.4367	8.8	6.0	0.1814
19	12.6	7.0	0.0807	9.8	5.0	0.0808
20	7.4	9.0	0.6000	1.8	8.0	0.9853
21	9.4	4.0	0.0512	6.5	4.0	0.1626
22	11.6	8.0	0.1691	7.2	8.0	0.5162
23	10.6	7.0	0.1539	8.4	7.0	0.2954
24	13.0	8.0	0.1097	8.1	9.0	0.5295
25	2.9	7.0	0.8977	16.6	7.0	0.0201
26	6.9	7.0	0.4418	19.5	8.0	0.0125
27	14.3	8.0	0.0734	10.9	8.0	0.2044
28	8.1	9.0	0.5207	6.6	6.0	0.3638
29	7.6	9.0	0.5739	12.9	8.0	0.1160
30	8.9	8.0	0.3530	22.3	9.0	0.0180
31	11.3	8.0	0.1828	11.1	8.0	0.1955
32	9.7	8.0	0.2836	16.9	8.0	0.0309
33	7.1	9.0	0.6249	5.2	7.2	0.6380
34	14.9	8.0	0.0614	2.3	7.0	0.9436
35	8.5	9.0	0.4833	7.7	8.0	0.4630
36	2.7	8.0	0.9516	4.8	8.0	0.7807
37	3.2	7.0	0.8709	4.1	7.0	0.7697
38	4.4	8.0	0.8157	9.2	8.0	0.3280
39	7.7	8.0	0.4653	13.1	8.0	0.1081
40	12.7	8.0	0.1228	5.8	8.0	0.6739
Whole Test	353.7	304.0	0.0262	331.6	286	0.0328

Note: DF or DF* is same as degrees of freedom, Prob-value or Prob-value* is same as probability value. Chi-square, DF, and Prob-value relate to item and whole test fit statistics for Math Placement Test form A, while Chi-square*, DF*, and Prob-value* relate to item and whole test fit statistics for Math Placement Test form G

Table 8 . The Discrimination and Difficulty parameter of Math Placement Test form A and G in a Comparative Matrix.

Items	a	a^*	b	b^*
1	0.749	0.405	-2.085	-1.327
2	0.817	0.876	-3.254	-2.997
3	1.592	1.377	0.235	0.390
4	1.376	1.301	0.272	0.320
5	0.715	0.584	-2.271	-3.610
6	0.839	0.606	-0.855	-1.701
7	0.879	0.416	0.346	-2.431
8	0.354	0.340	0.274	-0.324
9	1.455	1.601	0.130	-0.137
10	0.799	1.712	-1.789	-1.178
11	1.078	1.283	0.998	0.790
12	1.265	1.423	0.127	0.115
13	2.166	1.849	0.119	-0.275
14	0.608	0.334	0.400	2.703
15	0.605	0.648	-0.755	-0.988
16	0.594	0.657	1.609	0.656
17	1.471	1.586	-0.139	-0.176
18	1.868	1.472	0.538	0.146
19	1.251	1.561	-0.109	0.004
20	0.785	0.978	1.144	1.020
21	2.490	2.340	-0.091	0.023
22	1.414	0.994	0.560	1.139
23	1.789	1.346	0.224	0.226
24	2.062	0.500	1.935	1.857
25	2.068	1.348	0.662	0.657
26	1.597	0.957	0.918	1.624
27	1.252	0.858	1.965	1.698
28	0.883	1.721	0.600	0.423
29	0.504	0.704	-0.202	1.278
30	1.478	0.525	1.637	0.869
31	1.590	0.784	1.720	1.942
32	1.063	0.591	2.875	1.863
33	1.155	1.321	1.138	1.096
34	1.158	0.709	2.933	2.677
35	0.560	0.824	3.357	1.466
36	1.259	0.764	2.028	2.012
37	1.371	0.458	3.264	4.265
38	1.528	0.987	0.666	-0.349
39	1.735	0.420	1.909	3.294
40	0.909	0.830	2.497	2.097

Note: a represents discrimination index for placement test form A, a^* represents discrimination index for placement test form G, b represents item difficulty for placement test form A and b^* represents item difficulty for placement test form G

REFERENCES

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the Bulletin. Psychological Bulletin, 112, 400-404

BILOG (1990). Scientific Software International, Chicago, Il.

Birnbaum, A.(1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Norvick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 33, 205-33.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H.(1991). Fundamentals of Item Response Theory. Newbury Park, CA: SAGE Publications.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter models. Educational Measurement: Issues and Practice, 8, 35-41.

Hulin, L. C., Drasgow, F., & Parsons, C. K.(1983). Item Response Theory. Application to Psychological Measurement. Homewood, Illinois: Dorsey Professional Series.

Lord, F. M. (1952). A theory of test scores. Psychometric Monograph. No 7.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Erlbaum.

Lord, F. M. & Novick, M. R. (1968) Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

MicroFACT(1995) Assessment Systems Corporation, St Paul, MN.

Mislevy R. J. & Bock, R. D. (1990). Item Analysis and Test Scoring with Binary Logistics Models. Scientific Software, Inc.

Novick, R. M. & Lewis C. (1967). Coefficient alpha and the reliability of composite measurements. Psychometrika, 32, 1-13

Rasch, G.(1966). An item analysis which takes individual differences into account. The British Journal of Mathematical and Statistical Psychology, 19, 49-57.

Rasch, G.(1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4(3) 207-230

Statistical Package for Social Sciences (SPSS, Version 8.0). Chicago, IL.